Robust scaling in fusion science: Case study for the L-H power threshold

G Verdoolaege[1,2] and J-M Noterdaeme[1,3]

[1]Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium
[2]Laboratoire de Physique des Plasmas de l'ERM – Laboratorium voor Plasmafysica van de KMS (LPP–ERM/KMS), Ecole Royale Militaire – Koninklijke Militaire School, B-1000 Brussels, Belgium
[3]Max Planck Institute for Plasma Physics, Boltzmannstr. 2, 85748 Garching, Germany

E-mail: geert.verdoolaege@ugent.be

Abstract. In regression analysis for deriving scaling laws in the context of fusion studies, usually standard regression methods have been applied, of which ordinary least squares (OLS) is the most popular. However, concerns have been raised with respect to several assumptions underlying OLS in its application to fusion data. More sophisticated statistical techniques are available, but they are not widely used in the fusion community and, moreover, the predictions by scaling laws may vary significantly depending on the particular regression technique. Therefore we have developed a new regression method, which we call *geodesic least squares* regression (GLS), that is robust in the presence of significant uncertainty on both the data and the regression model. The method is based on probabilistic modeling of all variables involved in the scaling expression, using adequate probability distributions and a natural similarity measure between them (geodesic distance). In this work we revisit the scaling law for the power threshold for the L-to-H transition in tokamaks, using data from the multi-machine ITPA databases. Depending on the model assumptions, OLS can yield different predictions of the power threshold for ITER. In contrast, GLS regression delivers consistent results. Consequently, given the ubiquity and importance of scaling laws and parametric dependence studies in fusion research, GLS regression is proposed as a robust and easily implemented alternative to classic regression techniques.

PACS numbers: 02.50.Cw, 02.40.Ky, 52.55.Dy

1. Introduction

Statistical regression methods play a very important role in fusion data analysis, as one of the main activities in making physics inferences from data in fusion experiments. On the one hand, regression analysis is used for fitting deterministic relations reflecting physical dependencies between plasma variables. This is an essential instrument for evaluating theoretical predictions and for supporting theory building. On the other hand, scaling laws fitted to multi-machine databases provide design guidelines for future devices, by extrapolating key quantities along a regression line. Important examples are the energy confinement time and the threshold for the power required for the transition from the L-mode to the H-mode in tokamaks [1].

Ordinary least squares (OLS) regression is the statistical workhorse that is employed for these purposes in the vast majority of cases, primarily owing to its ease of implementation and availability in any software package for statistical regression. Various assumptions underly OLS and, while in many simple cases these approximate the true situation relatively well, fusion data can have quite rich distributional properties with complex nonlinear relations among variables. As a result, OLS may yield unreliable estimates for the regression parameters, adversely affecting theory building and predictions from scaling laws [2].

Putting the issue in the right perspective, one might observe that great efforts go into the careful design and operation of fusion diagnostics, and sophisticated theoretical models and modeling codes are developed. Therefore, to link these activities it is equally mandatory to employ state-of-the-art techniques from probability theory, statistics and machine learning for validating, processing and analyzing the data. As far as regression analysis is concerned, this is already relatively well accepted in many scientific fields that rely heavily on regression and scaling, such as astronomy, biology and ecology. In fusion science, however, this practice is considerably less widely spread. While in some cases OLS regression is certainly adequate, in many more complex situations OLS is not valid and will produce simply wrong results.

Unfortunately, the complexities of fusion data are very diverse and, while regression methods have been developed to address specific violations of the OLS assumptions, this covers an entire domain in statistics and probability theory. Each method requires its proper techniques and the literature is vast, so for non-experts it can be difficult to enter into the applications. Moreover, designing a robust regression model can be a complicated matter, requiring many decisions tailored to the problem at hand or rather *ad hoc*, which may or may not alter the results, possibly even leading to a loss of precision. In such cases, a more structural solution is desirable.

For these reasons we have developed a new regression method, called *geodesic least squares* regression (GLS), which is based on simple and straightforward principles and yet is sufficiently flexible and robust to address the complexities of fusion data in a unified way. The primary aims of this paper are to point out some of the dangers of an overly simple regression methodology and to present GLS as an alternative that

is well-grounded in probability theory, yet easily implemented by practitioners in the field, not necessarily with a background in probability theory. We introduce the GLS method and we discuss some of its advantages over OLS regression, as well as maximum *a posteriori* estimation (MAP), which is a well-known Bayesian method. Next, we present several regression experiments using synthetically generated data and we show the enhanced robustness of the method, relative to OLS and MAP, against outliers and model uncertainty originating from a logarithmic transformation of the data. These experiments are inspired by our case study in this paper, which is the well-known scaling of the L-H power threshold in the high-density branch. Reliable predictions of the L-H power threshold as well as the details of its parametric dependence are of great practical value for development of ITER plasma scenarios. Advanced (non-power-law) regression functions and determination of an optimal set of predictor variables, have been the subject of recent investigations, in relation to the L-H power threshold [3, 4]. In the present study, however, we concentrate on demonstrating the performance of GLS regression for scaling of the L-H power threshold. We base this on the standard regression model and the usual set of variables [5]. After presenting the results of the experiments with synthetic data, we provide a demonstration of the failure of OLS regression in consistently estimating and extrapolating the power threshold scaling law. We show that the results obtained by GLS are more robust, in comparison with both OLS and MAP, across different regression models and versions of the database.

The remainder of the paper is structured as follows. We start in Section 2 by introducing the principles of GLS regression and its advantages over OLS and MAP. A brief overview of the background related to information geometry is provided here, which is required for the description of the methodology. We introduce our case study related to power threshold scaling in Section 3, together with some general information about the multi-machine databases. The numerical experiments on synthetic data are discussed in Section 4, while Section 5 is devoted to the experiments and discussion concerning the power threshold scaling law, using the actual data from the international multi-machine databases. Finally, conclusions are drawn in Section 6.

2. Geodesic least squares regression

The necessity of an advanced approach to regression analysis when dealing with data from fusion experiments, fundamentally originates in the complexity of the physical system (the fusion plasma) and the measurement system (diagnostics in a hostile environment). This results in uncertainty on physical models and data models, which has to be addressed by means of dedicated statistical techniques. We start the presentation of GLS regression by briefly addressing the various complexities of fusion data, in relation to regression analysis. We will consider here so-called multiple regression, involving several predictor variables x_j $(j = 1, \ldots, m)$ and a single response variable y. Our point of view regarding probability theory is Bayesian (although in its present form GLS regression is not yet a fully Bayesian method; see below).

2.1. Fusion data characteristics

One of the main premises of the GLS regression method is motivated by the often strongly stochastic character of fusion data. Put simply, stochastic uncertainty is caused by measurement noise and plasma fluctuations, and this may result in significant error bars and non-Gaussian distributions. Consequently, it makes little sense to characterize the physical quantity of interest merely by a single measurement value. Instead, one could perform a series of repeated measurements and provide a summary of the distribution underlying these measurements. In case the distribution of a scalar quantity displays Gaussian characteristics, one could then mention estimates for its mean and standard deviation. For more general distributions it might be feasible to estimate higher-order moments. Another way to estimate probability models in fusion science is by calculating the distribution from a raw data set using Bayesian probability theory [6, 7, 8].

The key point is that the moments of the distribution of a plasma quantity, or, even more accurate, the distribution itself, contains important information about (our knowledge of) that quantity, beyond a single value or even a sample average. This realization, that a more complete and rich source of information lies in the probability distribution of a quantity of interest, is at the heart of GLS regression [9, 10, 11, 12, 13].

Naturally, in regression analysis not only the response variable but also the predictor variables are affected by noise. It is important to note, however, that classic OLS regression is based on the assumption of error-free predictor variables (infinite measurement precision). In many applications this can be seen as a relatively good approximation, because often the predictor variables have a significantly lower measurement uncertainty, or they can be better controlled, compared to the dependent variable. But in fusion applications the approximation can be too crude, and one needs to account for stochasticity of the predictor variables too [11]. In fact, this is one of the properties of fusion data that conflicts most often with the assumptions of OLS regression.

In frequentist statistics, uncertainty on all variables is handled by so-called 'errors-in-variables models', see for instance [14]. One of the main reasons why this problem is more difficult than the simple case of error-free predictor variables, is that the 'true' values of the predictor variables are unknown. Hence additional unknowns are introduced for every data point. Through errors-in-variables models, various remedies have been proposed to deal with this indeterminacy. Unfortunately, many of these have a rather *ad hoc* character and depend on additional assumptions. In contrast, a simple structural Bayesian solution has been outlined in [15, 16, 17], adequately addressing the issue of non-negligible stochastic uncertainty on the predictor variables. Our GLS method is partly inspired by this Bayesian solution to regression analysis, in the presence of errors on all variables.

On top of stochastic uncertainty on the measurements, there could be systematic measurement uncertainty. In a Bayesian context systematic uncertainty can be modeled

by appropriate nuisance parameters, but we will not specifically address that issue here. Furthermore, there could be uncertainty in the regression model, which in turn can be subdivided in two components. The first, deterministic component of the regression model is the functional form that is assumed to model the deterministic dependencies of the response variable on the predictor variables. The second, stochastic component concerns the model for the probability distribution that is assumed to describe the noise on the data. It may happen that the true regression function, the relevant set of predictor variables or the true distribution of the data, are quite different from what is suggested by the model assumptions. For instance, one particularly critical issue in deriving fusion scaling laws is the practice of converting the power-law scaling into a linear regression problem by transforming the variables to logarithmic space [1]. Despite this being a wide-spread habit in many areas of science, it is a well-known fact in probability theory that the logarithm (heavily) distorts the distribution of the data [2, 18]. It may seem that taking the logarithm leads to a simplified problem with the additional 'advantage' that extrapolation from the scaling law is straightforward, towards a point not far off the main data cloud on the logarithmic scale. In reality it is difficult to draw reliable conclusions from such an analysis and we will demonstrate below that a logarithmic transformation should be avoided.

A further complication that is not covered by standard OLS is heteroscedasticity, i.e. the fact that not all measurements of a certain quantity are equally noisy. Particularly in the case of multi-machine scaling laws this assumption is not fulfilled, as the same quantity is measured by different diagnostics on different machines. In addition, there may be statistical correlations between plasma parameters and the distributions of the variables involved can be non-Gaussian. Gaussianity (of the response variable) is not strictly an assumption of OLS regression (although zero skewness is), but it is often assumed to obtain tractable distributional properties of the estimated parameters. However, non-Gaussian or skewed distributions also occur frequently in fusion data, either when fitting directly to the data, or when calculating the distribution of derived quantities from the raw data using Bayesian probability theory. Finally, OLS regression can yield inaccurate results in the presence of atypical observations (outliers) or in the event of insufficient linear independence among the predictor variables (near-collinearity).

In a particular case where one or multiple assumptions of OLS are questionable, GLS regression can be used to address each of these issues in a single integrated framework. In the form that will be presented here, GLS still requires the data analyst to formulate the deterministic and stochastic components of the regression model (although non-parametric extensions could be envisaged), but the key difference with most existing regression techniques is that the dependence of the results on the model assumptions is greatly reduced. This is a very useful feature for fusion data analysis. Specifically, on the one hand, GLS considers the *modeled distribution* of the response variable that would be expected if all assumptions of the regression model were true (both deterministic and stochastic). This includes modeling of the uncertainty on the

predictor variables. On the other hand, an estimate is made of the *observed distribution* derived from the actual measurements of the response variable, with minimal additional assumptions. As opposed to OLS, and, indeed, most existing regression methods, GLS regression does not require both distributions to be the same, but rather it minimizes the difference between them. More precisely, GLS minimizes the *geodesic distance* between the distributions, which is a natural and mathematically well-founded similarity measure between probability distributions [10, 19, 20]. As such, GLS does not rigorously impose the assumptions of the regression model on the data, instead leaving sufficient flexibility to allow deviations from the chosen regression model.

Finally, any physics knowledge that may help to estimate the regression parameters, or physics-based constraints on the parameters, can be taken into account within the GLS formalism. For example, such information may guide the choice of the regression model. In addition, the geodesic-based regression method is presently based on optimization, which can be performed under known constraints. Moreover, in future developments the new method will be embedded in the Bayesian formalism, at which point it will become possible to encode physics knowledge into the prior distribution. However, it is important to note that also from a Bayesian point of view, the geodesic-based regression is fundamentally different from established techniques, and more general.

2.2. GLS methodology

The new GLS regression method presented here is a straightforward generalization of OLS and the basic principles have been discussed earlier in [9, 11, 12, 13]. Here, we provide a slightly more general introduction to GLS, by extending the classic multiple linear regression problem.

2.2.1. Standard regression analysis A parametric multiple regression problem can be formulated through a model function f that is nonlinear in general. f has some flexibility that is determined by p parameters β_k ($k = 1, \ldots, p$) (e.g. regression coefficients in linear regression). Let us suppose for now that all measurements are infinitely precise, i.e. there is zero noise on all variables. Given n realizations (measurement values) ξ_{ij} for each of m predictor variables ξ_j ($i = 1, \ldots, n$, $j = 1, \ldots, m$), the regression function produces n values η_i for the response variable η:

$$\eta_i = f(\xi_{i1}, \ldots, \xi_{im}, \beta_1, \ldots, \beta_p) \equiv f(\{\xi_{ij}\}, \{\beta_k\}), \qquad \forall i, \tag{1}$$

where we have introduced the notation $\{\xi_{ij}\}$ for the set of all ξ_{ij}, and likewise for $\{\beta_k\}$. In reality, all variables can be affected by noise, which for now we assume to be of a Gaussian nature, although this could be any distribution. Hence, all we have is a series of noisy measurements x_{ij} and y_i for the predictor and response variables x_j, resp. y:

$$\begin{aligned} y_i &= \eta_i + \epsilon_y, & \epsilon_y &\sim \mathcal{N}\left(0, \sigma_y^2\right), \\ x_{ij} &= \xi_{ij} + \epsilon_{x,j}, & \epsilon_{x,j} &\sim \mathcal{N}\left(0, \sigma_{x,j}^2\right). \end{aligned}$$

Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the normal probability distribution with mean μ and standard deviation σ. Note that we explicitly allow for the challenging case of non-negligible uncertainty on the predictor variables, which may be different for different variables. Also, we have described the simplified case of homoscedasticity: all measurements of a particular variable are assumed to be sampled from the same distribution. This can easily be generalized, however.

The principle of OLS regression is to find the parameter estimates $\hat{\beta}_k$ that minimize the sum of squared differences between the observations y_i of the response variable and their respective modeled values through the function f:

$$\{\hat{\beta}_k\} = \arg\min_{\{\beta_k\}} \sum_{i=1}^{n} \Big[y_i - f(\{x_{ij}\}, \{\beta_k\}) \Big]^2. \tag{2}$$

However, it is known that this produces unreliable results if the x_j are affected by noise that is not negligible compared to the noise on y [2, 14, 16]. A way around this is to consider the more general maximum likelihood method (ML). This involves maximizing the probability distribution of the response variable *conditional on* the predictor variables. Continuing with the case of a normal distribution on the response variable, this comes down to the following optimization problem (the σ_{mod} notation is explained below):

$$\{\hat{\beta}_k\} = \arg\max_{\{\beta_k\}} \frac{1}{\sqrt{2\pi}\sigma_{\text{mod}}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{\Big[y_i - f(\{x_{ij}\}, \{\beta_k\}) \Big]^2}{\sigma_{\text{mod}}^2} \right\}. \tag{3}$$

Here, we have assumed that the samples y_i have been realized in an independent way, that the variables x_j are mutually independent and that their realizations x_{ij} have also been drawn in an independent way. All these assumptions can be generalized. The distribution in (3) is called the likelihood of the model. The standard deviation σ_{obs} in general describes uncertainty on the response *and* the predictor variables. Indeed, the uncertainty on the predictor variables propagates through the function f and in (3) we have assumed that the result $f(\{x_{ij}\}, \{\beta_k\})$ is still Gaussian (therefore so is $y_i - f(\{x_{ij}\}, \{\beta_k\})$), or can be satisfactorily approximated by a Gaussian. However, in a more general setting, particularly for strongly nonlinear functions f, it should be noted that $f(\{x_{ij}\}, \{\beta_k\})$ may very well have a distinctly non-Gaussian shape. In that case there is a problem with one of our premises, as then it makes little sense to model the response variable by a normal distribution. We do not treat the full complexity of this issue here and instead focus on the case where the Gaussianity of y and $f(\{x_{ij}\}, \{\beta_k\})$ is a reasonable approximation. We then need to find a good approximation for the standard deviation σ_{mod} in (3). In addition, it is clear that, in the case of a Gaussian error distribution and neglecting the error bars on the x_j, the optimization in (3) is equivalent to OLS in (2).

We furthermore note that the maximum likelihood method can be extended to the Bayesian framework, by multiplying the likelihood distribution by appropriate prior distributions for the regression parameters. Maximization of the resulting posterior

distribution then leads to the maximum *a posteriori* method (MAP), which, together with OLS, we will use in the experimental sections for comparison with GLS. Just like maximum likelihood, MAP can take into account the uncertainty on the predictor variables—a quality which they share with GLS.

One particularly convenient model is the linear one:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \epsilon_y, & \epsilon_y &\sim \mathcal{N}\left(0, \sigma_y^2\right), \\ x_{ij} &= \xi_{ij} + \epsilon_{x,j}, & \epsilon_{x,j} &\sim \mathcal{N}\left(0, \sigma_{x,j}^2\right). \end{aligned}$$

Indeed, through marginalization of (integration over) the unknown 'true' variables ξ_j, it can be shown that the conditional distribution of y, given a measurement x_{ij}, is still Gaussian [15, 16]:

$$\begin{aligned} p_{\mathrm{mod}}(y|\{x_{ij}\},\{\beta_k\}) &= \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{mod}}} \exp\left[-\frac{(y-\mu_{\mathrm{mod},i})^2}{2\sigma_{\mathrm{mod}}^2}\right], \\ & j = 1,\ldots,m, \quad k = 1,\ldots,p. \end{aligned} \tag{4}$$

Here, we have defined

$$\mu_{\mathrm{mod},i} \equiv \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im}, \tag{5}$$

$$\sigma_{\mathrm{mod}}^2 \equiv \sigma_y^2 + \beta_1^2 \sigma_{x,1}^2 + \ldots + \beta_m^2 \sigma_{x,m}^2. \tag{6}$$

This could also have been obtained from standard Gaussian error propagation rules (with the same underlying assumptions). From now on, we furthermore suppose that the standard deviations $\sigma_{x,j}$ and σ_y are known. For instance, they could be defined as the error bars on the corresponding measurements. Again, this is an assumption that can be relaxed. We will call p_{mod} in (4) the *modeled distribution* of y, conditional on the measured values of the predictor variables.

2.2.2. Extending to GLS We now describe the key difference of GLS regression compared to existing methods. In classic regression, as described above, the goodness of the estimates of the model parameters β_k is measured purely by the likelihood of the data $\{y_i\}$ under the proposed regression model. In other words, it is assumed that the data points y_i are samples from the likelihood. Any deviations of either the distribution of the data, or the deterministic regression function from the proposed model, are likely to cause unreliable estimates of the model parameters. For this reason we introduce additional flexibility in the model, in that we will allow the true distribution of the data to deviate from the proposed model. This extra flexibility is expected to allow for model inaccuracies or model deviations.

In this simple example we will still assume that in reality the data have a normal distribution. The added flexibility is realized by explicitly modeling the standard deviation of the response variable y by an extra parameter σ_{obs}. It is this parameter that is expected to capture deviations from the model assumptions. The mean of this Gaussian, which we will call the *observed distribution* p_{obs} of y, is taken at each of the

individual data points. That is, the observed distribution of y, given the measurement point y_i, is given as follows:

$$p_{\rm obs}(y|y_i,\sigma_{\rm obs}) = \frac{1}{\sqrt{2\pi}\sigma_{\rm obs}} \exp\left[-\frac{(y-y_i)^2}{2\sigma_{\rm obs}^2}\right]. \tag{7}$$

Note that we have again assumed homoscedasticity, as $\sigma_{\rm obs}$ is the same for all measurements. This provides another opportunity for generalization of the method.

The aim of GLS is now to estimate the regression parameters—in the present example the β_k—by minimizing the difference (maximizing the similarity) between the modeled and the observed distribution. The question remains which similarity measure, or measure of distance, to use between the two distributions. For this, we employ a natural distance measure defined within a geometric approach to probability theory, called *information geometry* [21].

2.2.3. The geometry of probability theory In information geometry, a probability density family is interpreted as a (Riemannian) differentiable manifold (multidimensional surface). A point on the manifold corresponds to a specific probability density function (PDF) within the family and the family parameters provide a coordinate system on the manifold. The Fisher information, a well-known concept in statistics, plays the role of a unique metric tensor (Fisher-Rao metric). For a probability model $p(\boldsymbol{x}|\boldsymbol{\theta})$ describing a vector $\boldsymbol{x}$, parameterized by an m-dimensional vector $\boldsymbol{\theta}$ with components θ_i $(i = 1,\ldots,p)$, the entries g_{ij} of the Fisher information matrix are the following:

$$g_{ij}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\boldsymbol{x}|\boldsymbol{\theta})\right], \quad i,j = 1\ldots p.$$

Here, $\mathbb{E}$ signifies the expectation. Equipped with the Fisher-Rao metric one can calculate geodesics and the *Rao geodesic distance* (GD) between two points on the manifold. This sequence of steps is schematized in Figure 1. We do not go further into the mathematical details, which can be found in [9], [20] and [21]. Suffice it to mention here that the GD between two univariate normal distributions $p_1(x|\mu_1,\sigma_1)$ and $p_2(x|\mu_2,\sigma_2)$ is given by the following closed-form expression [22]:

$$\begin{aligned} {\rm GD}(p_1,p_2) &= \sqrt{2}\ln\frac{1+\delta}{1-\delta} = 2\sqrt{2}\tanh^{-1}\delta, \\ \delta &\equiv \left[\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}\right]^{1/2}. \end{aligned} \tag{8}$$

One could argue that a more simple distance measure between PDFs may be obtained by calculating the Euclidean distance between their respective parameters. For instance, the Euclidean distance ${\rm ED}(p_1,p_2)$ between two normal distributions $p_1 = \mathcal{N}(\mu_1,\sigma_1^2)$ and $p_2 = \mathcal{N}(\mu_2,\sigma_2^2)$ could be defined by

$${\rm ED}(p_1,p_2) \equiv \left[(\mu_1^2-\mu_2^2)+(\sigma_1^2-\sigma_2^2)\right]^{\frac{1}{2}}. \tag{9}$$

The problem is that his cannot be a suitable distance between distributions, for it does not respect the intrinsic geometry of the set of probability distributions from a certain

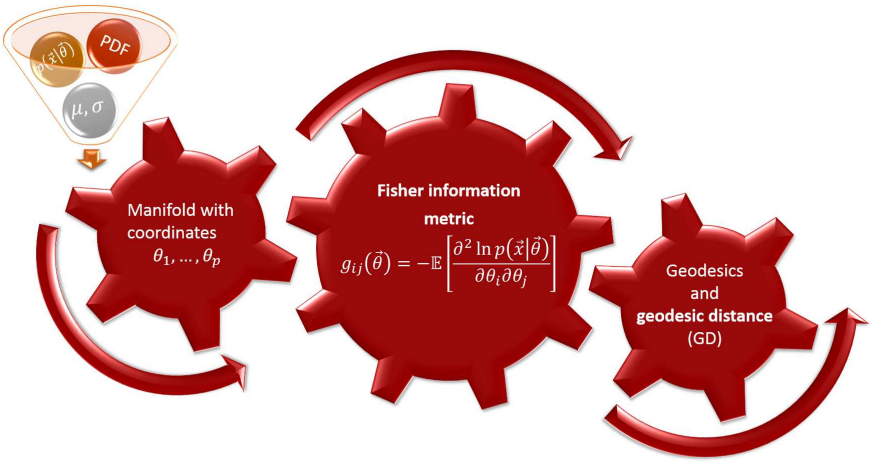


Figure 1: Schematic of the ingredients that enable the calculation of geodesic distances between probability distributions using information geometry.

family. We do not provide a rigorous proof of this statement here (see e.g. [21, 23]), but rather present an intuitively appealing argument by means of Figures 2(a) and (b). We consider two Gaussians with PDFs $p_1(x|4, 1.2)$ (i.e. $\mu = 4$, $\sigma = 1.2$) and $p_2(x|16, 1.5)$, drawn in Figure 2(a). In Figure 2(b) two Gaussians $p_3(x|4, 4.0)$ and $p_4(x|16, 5.0)$ are displayed, with the same respective means but larger standard deviations compared to the first case. Now, whereas p_1 and p_2 are easy to distinguish, the distributions p_3 and p_4 overlap to a much larger extent. This difference in the level of 'distinguishability' should, of course, be reflected in the distance between the distributions. That is, the distance between p_1 and p_2 should be larger than that between p_3 and p_4. Using the expression in (8) it can be seen that the GD fulfills this requirement: $\mathrm{GD}(p_1, p_2) = 5.3$ and $\mathrm{GD}(p_3, p_4) = 2.4$. On the contrary, the Euclidean distance between p_1 and p_2, calculated by means of (9), is 12.00, which is *smaller* than the Euclidean distance of 12.04 between p_3 and p_4. Also, as suggested by this example, the GD is more sensitive to differences in the standard deviations, compared to the Euclidean distance. Hence, the Euclidean distance does not properly take into account the intrinsically non-Euclidean character of probability distributions, exemplified in particular by the standard deviation in case of a normal distribution.

An instructive visualization of the two-dimensional surface of univariate Gaussians is provided by the *pseudosphere* (tractoid), pictured in Figure 2(c). Each point on this surface represents a normal distribution, with meridians representing lines of constant mean, while circles of latitude have a constant standard deviation. Although the pseudosphere exhibits some of the most important properties of the true geometry of normal distributions, it should be noted that it is still an imperfect model. Indeed, unlike the true Gaussian manifold, the pseudosphere is periodic in the mean μ and it is only valid for $\sigma > 1$. Nevertheless, it is interesting to visualize the geodesics between

the points corresponding to the distributions in Figures 2(a) and (b). One can visually check on the pseudosphere that the distance between p_3 and p_4 indeed has to be shorter than that between p_1 and p_2. A rescaled ('unwrapped') version of the pseudosphere is pictured in Figure 2(d), showing the geodesics more clearly (although distorted in the direction of μ).

2.2.4. GLS algorithm With the mathematical principles and tools discussed above, we are in a position to formulate the GLS algorithm. We first continue with the case of multiple linear regression and normal distributions. Assuming n independent realizations of the data set consisting of y_i and $\{x_{ij}\}$ $(i = 1, \ldots, n)$, the optimization task comes down to minimizing the GD between, on the one hand, a product of n observed distributions $p_{\text{obs}}(y|y_i, \sigma_{\text{obs}})$ and n modeled distributions $p_{\text{mod}}(y|\{x_{ij}\}, \{\beta_k\})$. It can easily be shown that the squared GD between two sets of products of distributions is given by the sum of squared GDs between the corresponding individual distributions [22]. Hence, the $p + 2$ parameters $\beta_0, \ldots, \beta_p, \sigma_{\text{obs}}$ are estimated by minimizing the following expression:

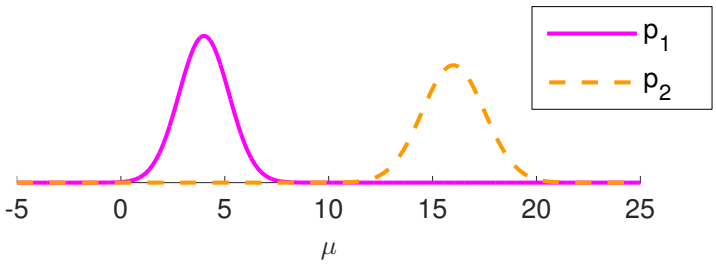
$$\begin{aligned}\{\hat{\beta}_k,\, \hat{\sigma}_{\text{obs}}\} &= \underset{\{\beta_k, \sigma_{\text{obs}}\}}{\arg\min}\, \text{GD}^2 \left[\prod_{i=1}^{n} p_{\text{obs}}(y|y_i, \sigma_{\text{obs}}), \prod_{i=1}^{n} p_{\text{mod}}(y|\{x_{ij}\}, \{\beta_k\})\right] \\ &= \underset{\{\beta_k, \sigma_{\text{obs}}\}}{\arg\min} \sum_{i=1}^{n} \text{GD}^2[p_{\text{obs}}(y|y_i, \sigma_{\text{obs}}), p_{\text{mod}}(y|\{x_{ij}\}, \{\beta_k\})] \\ &= \underset{\{\beta_k, \sigma_{\text{obs}}\}}{\arg\min} \sum_{i=1}^{n} \text{GD}^2[\mathcal{N}(y_i, \sigma_{\text{obs}}^2), \mathcal{N}(\mu_{\text{mod},i}, \sigma_{\text{mod}}^2)].\end{aligned}$$

As before, $\mu_{\text{mod},i}$ and σ_{mod} are given by (5) and (6), while the GD is calculated by means of (8). Thus, for a Gaussian distribution, GLS involves a comparison of not only the means, but also the standard deviations of the observed and modeled distributions. The observed distribution depends more purely on the data compared to the modeled distribution, and much less on the model assumptions. As a result, together with the added flexibility offered by the extra parameter σ_{obs}, GLS is less sensitive to incorrect model assumptions, as will become apparent in the experimental sections.

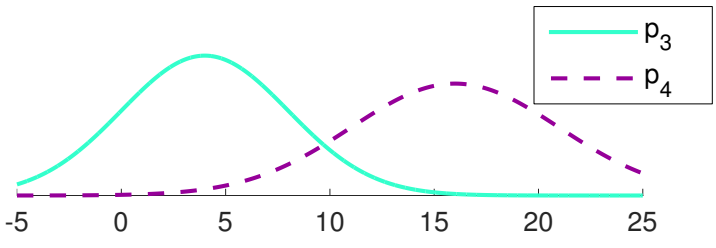
It is interesting to note that, if we would force $\sigma_{\text{obs}} \equiv \sigma_{\text{mod}}$, then the GD between the two Gaussian distributions $p_{\text{obs}}(y|y_i, \sigma_{\text{obs}})$ and $p_{\text{mod}}(y|\{x_{ij}\}, \{\beta_k\})$ would become [24]

$$\text{GD}(p_{\text{obs}}, p_{\text{mod}}) = \frac{|y_i - \mu_{\text{obs},i}|}{\sigma_{\text{obs}}}.$$

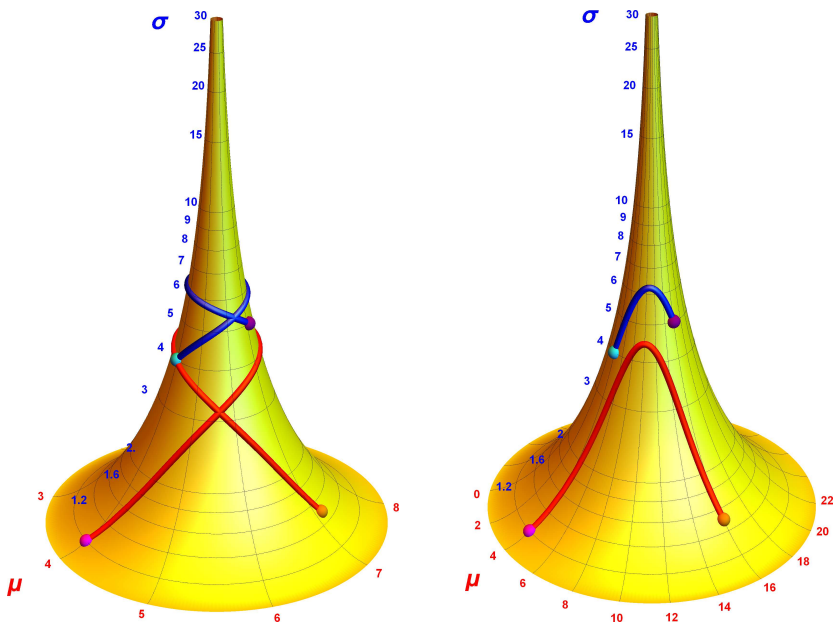
This is also called the Mahalanobis distance between the points y_i and $\mu_{\text{obs},i}$, assumed to be drawn from the same normal distribution with standard deviation σ_{obs}. But that would bring us right back to the maximum likelihood or maximum *a posteriori* method, for minimization of the sum of squared GDs is equivalent to maximization of the likelihood in (3). It is indeed desirable that GLS reduces to ML and MAP in the case of Gaussian distributions with identical standard deviations.



(a)



(b)



(c) (d)

Figure 2: In (a), the normal PDF p_1 (defined in the main text) is relatively far from p_2, compared to (b), wherein p_3 and p_4 have the same respective means, but are closer according to the GD. (c) The pseudosphere as a model for the univariate normal manifold. The parallels of the tractroid are lines of constant standard deviation σ, while the meridians (the tractrices) are lines of constant mean μ. This representation of the normal manifold is periodic in the μ-direction and a rescaled version (longer period along μ) is shown in (d). The distributions in (a) and the geodesics between them have been mapped on the surface of the pseudosphere in (c) and (d).

Having shown that OLS, ML and MAP are special cases of GLS, we stress again that GLS provides a solution to the robustness and stability issue of regression analysis that is fundamentally different from existing techniques. For instance, when the presence of outliers is suspected, common Bayesian approaches to robust regression analysis often use a heavy-tailed or mixture likelihood distribution, or adequate prior distributions are introduced for the regression parameters. However, it should be noted that similar measures can be taken in the case of GLS, although unnecessary here, but the resulting geodesic-based method would still be more general than the standard robust Bayesian approach. Moreover, the latter leads to a loss of precision when it turns out that, in reality, there are no outliers contaminating the data after all. In contrast, in the absence of contamination, GLS simply equalizes the values of the observed and modeled standard deviations, as will be shown in the experiments in Section 5. Furthermore, there are similarities of GLS with a class of methods known in the statistics literature as 'minimum distance estimation' (MDE) [25, 26]. However, there are also several differences, primarily in that GLS calculates the geodesic distance between each *individual* pair of modeled and observed distributions of the response variable, corresponding to an individual measurement point. As such, each individual data point acquires the status of a probability distribution in its own right. Consequently, GLS performs regression between probability distributions on a Riemannian probabilistic manifold. This is intrinsically different from classic regression methods, like OLS, ML and MAP, which operate in a flat Euclidean data space.

It was already mentioned that, in principle, the GLS procedure can be generalized to any deterministic regression function. With a view to the experiments in Sections 4 and 5, we now discuss the case of (nonlinear) power-law regression. In order to keep the computations tractable, we will assume that the uncertainty on the predictor variables is sufficiently small and the nonlinearity sufficiently weak in order to enable Gaussian error propagation. This approximation may be improved in future work. The power law relating the realizations x_{ij} of the predictor variables to the measurements y_i of the response variable, can be parameterized as follows, assuming additive Gaussian noise on all variables:

$$\begin{aligned} y_i &= \beta_0 x_{i1}^{\beta_1} \dots x_{im}^{\beta_m} + \epsilon_y, \qquad & \epsilon_y &\sim \mathcal{N}\left(0, \sigma_y^2\right), \\ x_{ij} &= \xi_{ij} + \epsilon_{x,j}, & \epsilon_{x,j} &\sim \mathcal{N}\left(0, \sigma_{x,j}^2\right). \end{aligned} \tag{10}$$

According to standard Gaussian error propagation laws, the modeled distribution, i.e. the distribution of the right-hand side in the expression for y_i in (10), can be approximated by a normal distribution with mean and standard deviation given by

$$\begin{aligned} \mu_{\mathrm{mod},i} &= \beta_0 x_{i1}^{\beta_1} \dots x_{im}^{\beta_m}, \\ \sigma_{\mathrm{mod},i}^2 &= \sigma_y^2 + \mu_{\mathrm{mod},i}^2 \left[\beta_1^2 \left(\frac{\sigma_{x,1}^2}{x_{i1}^2} \right)^2 + \dots + \beta_m^2 \left(\frac{\sigma_{x,m}^2}{x_{im}^2} \right)^2 \right]. \end{aligned} \tag{11}$$

Hence, the error bars depend on the measurements (heteroscedasticity). Nevertheless, we will introduce an approximation leading to constant error bars of measurements

originating from a single tokamak. This assumption may be relaxed in the future.

Finally, we still mention that, in the applications presented below, the minimization of the GD is a straightforward optimization problem that can be carried out by a generic algorithm. In the experiments we employed a classic active-set approach [27].

2.2.5. Credible intervals Presently, the GLS method does not directly offer confidence intervals on the estimated quantities. In this paper, the concept of a confidence interval—or more precisely: a credible interval—corresponds to the standard Bayesian definition of an interval wherein the true value of a stochastic variable is assumed to lie with a certain probability (e.g. 0.95). This is different from the confidence intervals mentioned in [5] and [28], where the possibility is considered that the deviation of the true parameter values from the estimated ones may be entirely systematic (although they are defined as standard errors, hence in fact they are of a stochastic nature). Then, the maximum deviations of the true power threshold from the predicted value are calculated, when the systematic errors on all parameters would reinforce the deviation. The method used in the present paper causes less extreme error bars, although the influence of systematic errors deserves to be further investigated. Future work will address the issue of credible intervals in more detail, but for now error estimates were delivered by Monte Carlo estimation in the case of synthetic data (Section 4) and by bootstrapping when using the real data (Section 5). Monte Carlo sampling simply refers to repeating the regression experiment several times, each time performing the sampling of the stochastic elements in the model for the synthetic data, such as the noise, synthetic outliers, etc. Then, the regression analysis is carried out on each of the data sets and Monte Carlo averages are calculated for the estimated coefficients. Bootstrapping, on the other hand, is a well-known technique in statistics, which involves creating a large number of artificial data sets from the measured data, by resampling with replacement [29]. The regression analysis is then carried out on each of the data sets and the mean and standard deviation, over all data sets, of each estimated regression parameter and of the predicted quantities (e.g. the L-H power threshold for ITER) are used as estimates of the parameter or prediction value and its error bar, respectively. We used 100 bootstrap samples in our experiments with real data. This scheme typically results in rather conservative error bars, which could possibly be narrowed down using more sophisticated methods.

3. Power threshold scaling and database

The most recent commonly cited multi-machine scaling for the power threshold was obtained by Martin *et al.*, in [5], using a selection of data from the International Tokamak Physics Activity (ITPA) multi-machine database for the L-H power threshold [30, 31, 32]. However, with the purpose of investigating the robustness of our estimates and predictions, we also performed the analysis on an older version of the database, which was used to construct a scaling law by Snipes *et al.* in [28].

Various criteria have been established to select in the databases measurements from ITER-like plasmas. These can be consulted in [30, 31, 5, 28] and we do not consider them here in detail. The data selected in [5], which we will refer to as the 'ITPA08' data set, consist of 1024 time slices originating from six devices: ASDEX Upgrade (AUG) (175 slices), Alcator C-Mod (C-Mod) (115), DIII-D (56), JET (562), JFT-2M (58) and JT-60U (58). The older data set described in [28], which we denote by 'ITPA02', contains 616 time slices from eight tokamaks: ASDEX (37 slices), ASDEX Upgrade (172), Alcator C-Mod (130), DIII-D (55), JET (118), JFT-2M (41), JT-60U (58) and PBXM (5). Compared to the ITPA02 data set, ITPA08 contains new and corrected time slices, and follows improved selection criteria, leading to a much improved data conditioning. These criteria include an ion $\boldsymbol{\nabla} B$ drift towards the X-point, deuterium plasmas and sufficiently high line-averaged electron density $\bar{n}_\mathrm{e}$—a regime where P_thr is seen to increase as a function of density. Furthermore, the power threshold is assumed to additionally depend on the vacuum toroidal magnetic field on the magnetic axis B_t and the plasma surface area S. The dependence is chosen according to the following power law:

$$P_\mathrm{thr} = \beta_0 \, \bar{n}_\mathrm{e}^{\beta_1} B_\mathrm{t}^{\beta_2} S^{\beta_3}, \tag{12}$$

where β_0, β_1, β_2 and β_3 are the regression parameters to be estimated. Here, P_thr is in MW, $\bar{n}_\mathrm{e}$ in 10^{20} m^{-3}, B_T in T and S in m^2. For the purpose of this paper, we will continue using this global scaling law, without going into details regarding dependencies of the power threshold suggested by recent physical models of the L-H transition.

The databases also contain some information regarding the error bars on the measurements. This is important for our purposes, because we need the error bars to calculate σ_mod in (6) or (11) (they define the σ_y and $\sigma_{x,j}$). In the database, relative errors are quoted that are expressed as percentages. Unfortunately, the precise definition of error bars quoted in fusion science is not always clear. Usually, an error bar represents an estimate by the diagnostician of the typical range in which the 'true' quantity can be expected, where the uncertainty is assumed to be caused by *both* stochastic and systematic effects. Moreover, often it is difficult to assess the probability that is covered by the stochastic component of the error. Since a detailed investigation of the uncertainty of the threshold data is beyond the scope of the present paper, we will assume that the error bars pertain to a stochastic uncertainty corresponding to a single standard deviation of a Gaussian distribution. For some derived quantities the error bars had to be calculated from the uncertainty on more fundamental measurements. In those cases we employed Gaussian error propagation rules to estimate the standard deviation on the derived quantities. For the case of the global H-mode confinement database, this strategy has been shown to provide reasonable information on the actual measurement error bars [10]. On average over all devices, the typical measurement error bars quoted in the ITPA02 database are estimated at 4% for $\bar{n}_\mathrm{e}$, 1% for B_t, 3% for S and 15% for P_thr [30, 31]. In the ITPA08 database, although the relative error bars are the same, the averages are somewhat different, primarily due to the different numbers

of contributed data samples for the various devices. The average error bars for the 2008 data are 6% for $\bar{n}_{\mathrm{e}}$, 1% for B_{t}, 4% for S and 14% for P_{thr}.

Nevertheless, it is important to mention that the uncertainty in the data used for power threshold scaling, when compared to the predictions of a simple power law regression model (often referred to as the distribution of the residuals), is not expected to be due only to the measurement uncertainty on the individual variables. Indeed, in regression analysis any deviation of a data point from the deterministic component of the model (e.g. the scaling law) is interpreted as due to 'random' effects or 'noise'. More precisely, the uncertainty can be described as being caused by mechanisms that are too complex to be modeled deterministically, or that are simply not the main subject of investigation of a specific analysis. Now, in the case of the multi-machine ITPA databases it is clear that, other than measurement error, there are additional sources of deviation of the data from the scaling law. This is mainly due to the simplicity of the model, which contains only a few global predictor variables, and variability between machines and between experiments. It is difficult to estimate this uncertainty, but we here provide upper bounds by means of the following calculation. First, the nonlinear scaling law was estimated using OLS (the reference), as explained in Section 5.2. Then, for a specific variable z (one of the predictor variables or the dependent variable) and for each data point, the *relative* difference was computed between the z-value of the data point itself, and the z-value of the projection of the data point on the hypersurface given by the scaling law, keeping the values of the other variables fixed. This difference can be interpreted as the deviation of the point from the theoretical scaling law, assuming the deviation is solely due to the variability of z. Finally, the standard deviation of these relative differences was taken and the procedure was repeated for every predictor variable and the dependent variable. The resulting standard deviations can be interpreted as upper bounds of the relative variability of each of the quantities around their 'theoretical' values given by the scaling law. When applying this procedure to the ITPA02 data set, we obtained much higher values than the estimated error bars due to measurement error alone, as seen in Table 1. On the other hand, using the same procedure on the ITPA08 database resulted in error bars that, for the predictor variables, are still higher than those expected purely on the basis of measurement error, yet drastically lower than the estimates obtained on the ITPA02 database; see Table 1. For P_{thr}, the procedure yields 5% using the ITPA08 data, which is even lower than the nominal 14% quoted in the database. This confirms the significantly better conditioning of the data in the 2008 database: the data cloud is less dispersed and more closely fits a deterministic relation. We end this discussion by stressing that the obtained error estimates are upper bounds, so they cannot be used as estimates of the actual data variability. For this reason, the capabilities of GLS (through σ_{obs}) to handle the larger uncertainty, relative to the uncertainty expected from measurement error alone, will turn out to be important.

Table 1: Estimates of the relative error bars (percentages) on the predictor and response variables in the ITPA02 and ITPA08 databases, relative to the power threshold scaling law estimated through nonlinear ordinary least squares regression.

	$\bar{n}_{\rm e}$	$B_{\rm t}$	S	$P_{\rm thr}$
ITPA02	39	31	28	38
ITPA08	7	7	5	5

4. Numerical simulations

We now present a series of experiments with synthetic data, in order to strengthen confidence in the proposed regression method. In these experiments, the deterministic part of the regression model is based on the real-world problem for the L-H power threshold in fusion plasmas, considered in Section 5. The values of the predictor variables are those in the database, from which the values of the response variable (normally $P_{\rm thr}$) are generated synthetically. We discuss three different experimental setups: linear regression with errors on the predictor and response variables, linear regression under the same circumstances but introducing some atypical observations (outliers) and linear regression carried out after a logarithmic transformation of a power law, with errors on all variables. These experiments complement earlier studies of the enhanced robustness of GLS against data outliers and logarithmic transformation using synthetic data [9, 13, 11].

4.1. Linear regression

In the first experiment, the data set was created as follows. First, an artificial linear regression law was put forward for a variable η, depending on the predictor variables $\bar{n}_{\rm e}$, $B_{\rm t}$ and S, which were introduced in the context of the power threshold scaling law in Section 3‡. In particular, we generated a number of realizations of the variable η from the following prescription:

$$\eta = \beta_0 + \beta_1 \bar{n}_{\rm e} + \beta_2 B_{\rm t} + \beta_3 S. \tag{13}$$

This was considered as the 'true' relation between the predictor and response variables, where, as mentioned above, the values of the predictor variables were chosen to be exactly those from the ITPA databases, which are normally used in the real power threshold scaling law. We performed the analysis both on the 2002 and 2008 versions of the database.

‡ We use the notation η for the response variable instead of $P_{\rm thr}$ because in this experiment η is generated artificially and therefore it is not necessarily related to the actual power threshold in fusion devices.

An entire range of data sets was created using the following values of the coefficients β_0, β_1, β_2 and β_3:

$$\begin{aligned} &\beta_0 = 1, 1.1, \ldots, 20, \\ &\beta_1, \beta_2, \beta_3 = 0.1, 0.2, \ldots, 2. \end{aligned} \tag{14}$$

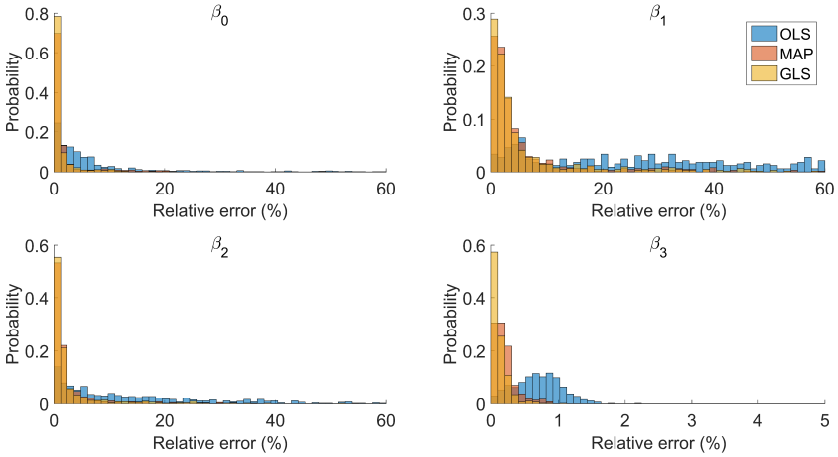
Thus, for each combination of values of β_0, β_1, β_2 and β_3, all 616 (1024) values of η were calculated according to (13), based on the values of $\bar{n}_\mathrm{e}$, B_t and S from the ITPA02 (ITPA08) database. The range of coefficient values in (14) was chosen to be representative for the values that are typically obtained from a regression analysis on the true scaling law (see Section 5). The exception is β_0, for which the range was chosen of roughly the same order as $\eta - \beta_0$ (much smaller values of β_0 would not be estimable in comparison with $\eta - \beta_0$).

Next, Gaussian noise was added to the predictor and response variables. The noise level was chosen according to the typical relative measurement errors in the ITPA02 database, i.e. (on average over all machines) 4% for $\bar{n}_\mathrm{e}$, resulting in a variable x_1, 1% for B_t (variable x_2), 3% for S (variable x_3) and 15% for the dependent variable (variable y, which is P_thr in the real-world regression problem). It should be stressed that, in the light of our comments in Section 3 regarding the variability of the predictor quantities, these are rather low noise levels. We further note that fixed relative noise levels lead to a different standard deviation for each measurement (heteroscedasticity). Accordingly, in implementing GLS a separate parameter describing the observed standard deviation should be introduced for each measurement point, in principle. As this would lead to unnecessary complications, we only defined one parameter $\sigma_{\mathrm{obs},\alpha}$ ($\alpha = 1, \ldots, N_\mathrm{t}$) for each of the N_t tokamaks contributing data to the database.

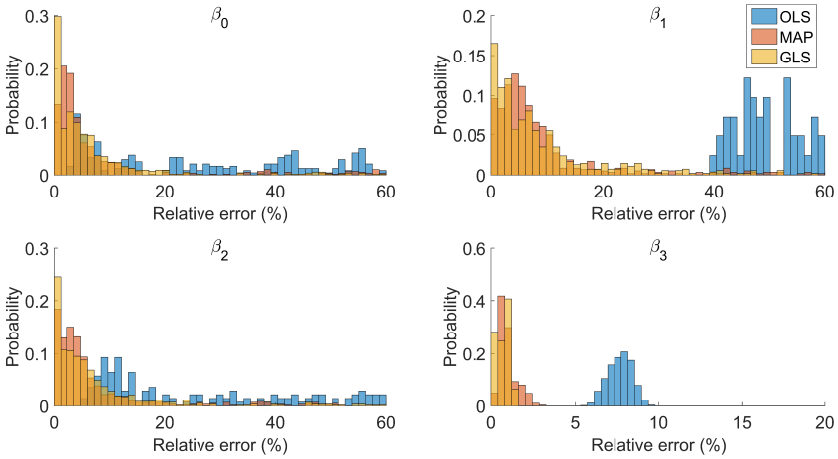
For each combination of coefficient values β_k ($k = 0, \ldots, 3$) taken from (14), 10 data sets were realized, each time performing the sampling of the noise. Finally, the regression analysis was carried out for every data set using OLS, MAP and GLS regression. As far as MAP is concerned, in the case of regression with uncertainty in predictor and response variables, special care has to be taken regarding the choice of maximally uninformative prior distributions for the parameters. We used the priors established in [16].

To report the results, for each choice of the β_k, the obtained estimates $\hat{\beta}_k$ were defined as the Monte Carlo average over the 10 data realizations. Next, histograms were created based on these averages for the estimated coefficients, specifically the normalized histograms of the relative difference $(\beta_k - \hat{\beta}_k)/\beta_k$ ($k = 0, \ldots, 3$), expressed as a percentage, between the true value β_k and the estimated value $\hat{\beta}_k$ of each regression parameter. The histograms of these percentage errors are shown in Figure 3(a), for the case of predictor values taken from the ITPA02 database, and in Figure 3(b) for the ITPA08 predictor values.

From the histograms it is clear that OLS does not perform well in estimating β_1 (coefficient of $\bar{n}_\mathrm{e}$) and β_2 (coefficient of B_t), with relative errors easily reaching 20-60%. In the case of the ITPA08 data, also the offset β_0 is poorly estimated by OLS. This classic method fails because it does not take into account the significant error bars on



(a)



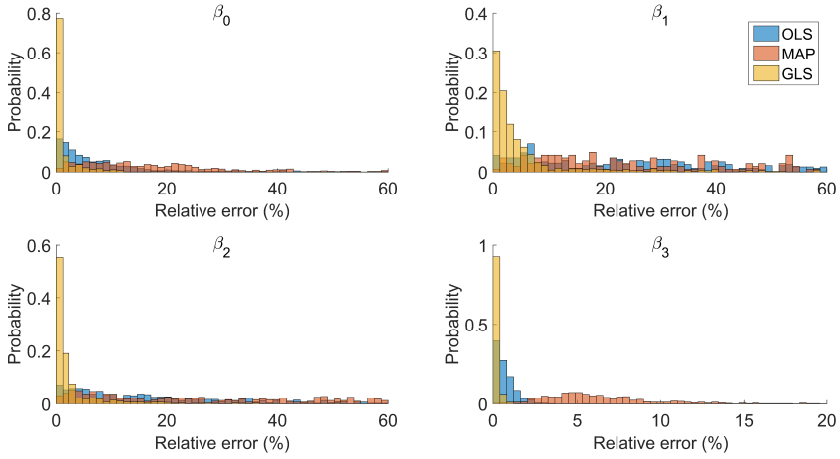
(b)

Figure 3: (a) Histograms of the relative error in estimating the regression coefficients β_k ($k = 0, \ldots, 3$) by means of OLS, MAP and GLS for an artificial linear regression problem. The values of the predictor variables were taken from the ITPA02 data. Horizontal axes represent the error in percent and vertical axes probability, normalized to 1. (b) Similar, for the ITPA08 data. Note the different scale on the abscissa for β_3, compared to (a).

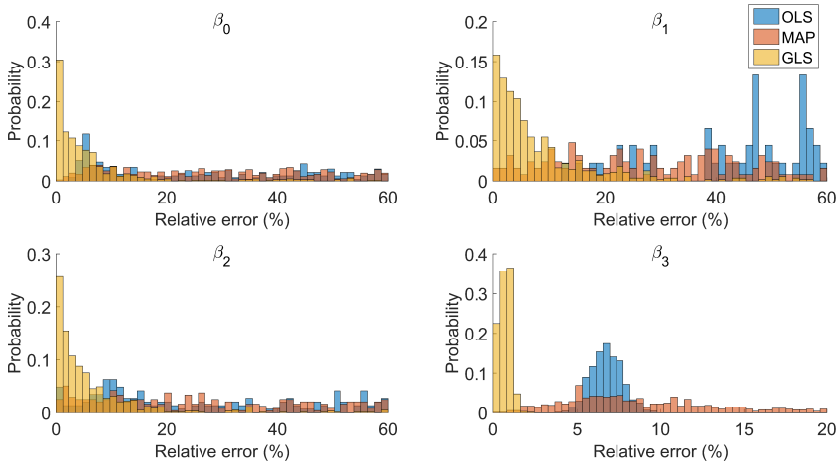
the predictor variables. The results of MAP and GLS are almost equally good, with only the estimates of β_1, associated to $\bar{n}_e$, occasionally off the true value by more than 20%. These are cases where, by chance, some unfavorable outliers where created by sampling from the noise distributions. In fact, the parameter that is overall most difficult to estimate turns out to be β_1. On the other hand, the coefficient of S is relatively stable.

4.2. Linear regression with outliers

In the next test we intended to examine the influence of outliers on the value of the dependent variable, deliberately introduced into the data set. The experimental setup



(a)



(b)

Figure 4: Histograms of the relative error in estimating the regression coefficients β_k ($k = 0, \ldots, 3$) by means of OLS, MAP and GLS for an artificial linear regression problem. Similar to Figure 3, but including 10 outliers.

was identical to that of the previous experiment, but, in addition, 10 outliers were created in each of the data sets. In particular, from the total of 616 points in each data set using ITPA02 data (1024 for the ITPA08 data), 10 points were randomly chosen and the associated value of the response variable y was multiplied with a factor F, where F was uniformly distributed between 1.5 and 2.5. Again, for each combination of coefficient values β_k ($k = 0, \ldots, 3$) taken from (14), 10 data sets were realized, each time performing the sampling of noise and outliers.

The results of carrying out the regression analysis by OLS, MAP and GLS on these synthetic data sets are shown in Figure 4. Now OLS and MAP perform much worse than GLS, both for the ITPA02 and ITPA08 data. In the case of GLS, the vast majority of relative errors is of the order of a few percent and certainly smaller than 20%. Again, the coefficient for $\bar{n}_\mathrm{e}$ is the most difficult to estimate, while the coefficient for S is more

stable.

The superior performance of GLS over OLS and MAP can be explained by the extra flexibility introduced through the observed standard deviation, which, in the case of outliers, is larger on average than the variability predicted by the model. Neither OLS, nor MAP possess this additional flexibility, instead forcing the unrealistic modeled standard deviation on the data. This is the primary asset of GLS, as explained in Section 2.

4.3. Log-linear regression

Finally, an experiment was carried out to point out the adverse effect of a logarithmic transformation, which is often used to transform a power-law regression model into a linear form. However, the logarithm alters the data distribution, which may lead to misguided inferences from OLS [2, 18]. Therefore the flexibility offered by GLS is expected to be beneficial in this case, as it allows the observed distribution to deviate from the modeled distribution.

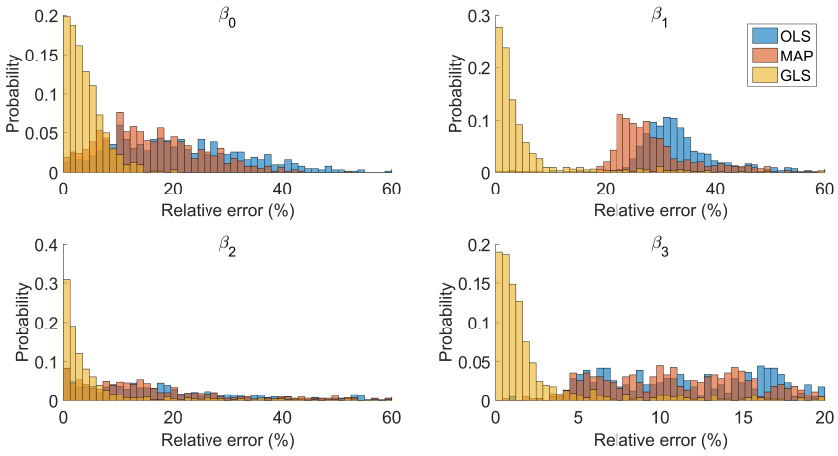
Again, the setup was very similar to the experiment in Section 4.1, however in the present case we started from a power-law deterministic model. In particular, the variable η was calculated for the same range of values of the parameters β_k as given in (14), but now according to a power law:

$$\eta = \beta_0 \bar{n}_\mathrm{e}^{\beta_1} B_\mathrm{t}^{\beta_2} S^{\beta_3}.$$

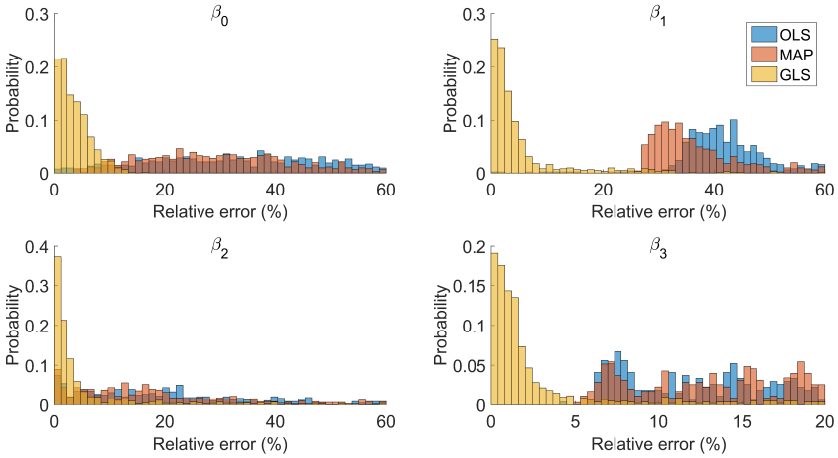
Then, Gaussian noise was added to all variables. However, when applying the relatively low noise levels used in Sections 4.1 and 4.2, only small differences were observed in the performance of GLS and MAP (see also the final test below). Therefore the noise levels for the predictor variables were augmented to (on average across all machines) 20% for $\bar{n}_\mathrm{e}$ (variable x_1), 5% for B_t (variable x_2) and 15% for S (variable x_3). The level for P_thr was kept at 15%, as before. This is still well within the maximum variability range that can be expected for the predictor variables in the ITPA02 database, as discussed in Section 3 (Table 1).

After adding the noise, all data were transformed to the logarithmic domain and 10 data sets were generated for each combination of regression coefficients. In GLS, the $\sigma_{\mathrm{obs},\alpha}$ now describe the observed standard deviations on the *logarithmic* power threshold. This, of course, corresponds to the relative errors on the power threshold itself.

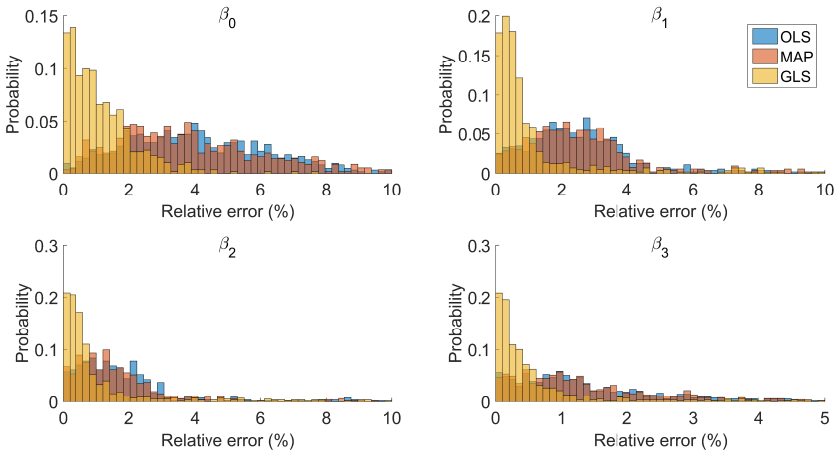
Subsequently, linear regression analysis was applied to each of the log-transformed data sets. The coefficient estimates, defined as the average over the 10 replications, were then compared among the various regression methods, as shown in Figure 5. Again, the normalized histograms of the relative error on the estimated parameters are displayed, showing the consistently better performance of GLS over OLS and MAP. For GLS, the errors on β_0 and β_1 are the largest, compared to those on β_2 and β_3, but the majority is still below 20%. As for β_0, the slightly inferior performance of GLS relative to the results with outliers in Section 4.2, is simply due to the fact that $\log \beta_0$ for the lowest values of β_0 is negligibly small compared to $\log \eta - \log \beta_0$.



(a)



(b)



(c)

Figure 5: Histograms of the relative error in estimating the regression coefficients β_k ($k = 0, \ldots, 3$) by means of OLS, MAP and GLS for an artificial log-linear regression problem. Similar to Figure 3, but with higher noise levels in (a) and (b). (c) Log-linear regression using predictor variables from the ITPA08 database, but with lower noise. Note the different scales on the abscissae, compared to (a) and (b).

The explanation for the better performance of GLS lies again in its added flexibility provided by the observed standard deviation. As a result, GLS is less restricted by the model assumptions, which, due to the logarithmic transformation, are incorrect.

We finally performed one more test based on the ITPA08 data, lowering the noise levels used in synthesizing the data to (on average over all devices) 5% for $\bar{n}_\mathrm{e}$, 5% for B_t, 3% for S and 3% for P_thr. These levels are somewhat lower than the maximum variability ranges seen in the ITPA08 database, listed in Table 1. As is to be expected, overall this does lead to substantially lower errors on the coefficient estimates using all regression techniques, but the trend remains the same: OLS and MAP perform significantly worse than GLS.

5. Power threshold scaling with GLS

We now come to the application of power threshold scaling using real-world data from the ITPA databases for all variables, including the response variable P_thr. We start with log-linear regression and then apply nonlinear regression analysis. It is important to note that we do not aim at a comprehensive database study here. Rather, we intend to demonstrate the power and consistency of GLS regression. The results of the experiments in this section are discussed in Section 5.3.

5.1. Log-linear scaling

We first followed the standard practice in transforming the power law (12) to the logarithmic scale to estimate the coefficients β_0, β_1, β_2 and β_3 via linear regression. To calculate σ_mod for each data point, we used the relative measurement error bars quoted in the database (typically 4% for $\bar{n}_\mathrm{e}$, 1% for B_t, 3% for S and 15% for P_thr). Considering the discussion in Section 3 regarding other sources of uncertainty, it is clear that the parameters $\sigma_{\mathrm{obs},\alpha}$ ($\alpha = 1, \ldots, N_\mathrm{t}$), describing the observed standard deviation in each of the N_t devices, will need to take into account other, 'unexpected' uncertainty sources, hence increasing the flexibility of the method.

The results of OLS, MAP and GLS regression on the ITPA02 data are given in Table 2. The predictions for ITER are also shown, for two typical densities (0.5 and 1.0×10^{20} m^{-3}). All estimates are accompanied by their 95% credible intervals obtained from 100 bootstrap samples. It is important to clearly state the interpretation of these intervals. For a given regression model and a given regression method, these error bars indicate the intervals in which the 'actual' values of the regression parameters lie with a probability of 0.95, based on the variability displayed by the data. This does not take into account, for instance, the possibility that the regression model might be suboptimal (e.g. not all predictor variables are taken into account), that the applied regression technique might be inadequate or that the data set is not representative of the true scaling law (in fact, these are issues that GLS aims to address). It explains why the regression results when using different methods and databases can be significantly

Table 2: Estimates of regression parameters β_k and predictions for ITER in log-transformed linear scaling of the H-mode threshold power using the ITPA02 data set. The bootstrap averages are given, as well as the 95% credible intervals (CI).

Method		β_0	β_1	β_2	β_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	Av.	0.051	0.49	0.87	0.84	38	53
	CI	±0.006	±0.07	±0.06	±0.04	±4	±8
MAP	Av.	0.045	0.57	0.87	0.90	46	68
	CI	±0.005	±0.08	±0.07	±0.04	±5	±9
GLS	Av.	0.043	0.66	0.80	0.95	48	76
	CI	±0.004	±0.07	±0.06	±0.03	±5	±9

Table 3: Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ on the logarithmic power threshold, expressed as percentage errors on P_{thr} itself, in the machines contributing to the ITPA02 data set, obtained using log-transformed linear scaling with GLS. The bootstrap averages are given, as well as the 95% credible intervals (CI).

	ASDEX	AUG	C-Mod	DIII-D	JET	JFT-2M	JT-60U	PBXM
Av. (%)	42	23	22	16	25	16	23	28
CI (%)	±5	±1	±1	±2	±2	±1	±2	±3

different, i.e. outside each other's credible intervals, as will be noted in the discussion section below. Also, we chose to mention only a single significant digit in the size of the credible intervals, in order to avoid the unrealistic impression of overly precise regression estimates.

The estimates by GLS of the parameters $\sigma_{\mathrm{obs},\alpha}$ (on $\log P_{\mathrm{thr}}$), including their credible intervals, for each of the devices contributing to the ITPA02 data, are given in Table 3. They have been expressed as relative errors on the bootstrap-averaged P_{thr}. The relative error on the power threshold lies around 20–30% for the various machines, except for ASDEX, where the uncertainty reaches a higher level of about 40%.

The outcome of similar calculations on the ITPA08 data set are presented in Tables 4 and 5.

5.2. Nonlinear scaling

Next, we show the results of nonlinear regression in the original data space, i.e. without logarithmic transformation. Whereas this prevents an analytic solution using OLS, the advantage is that the distribution of the data is left undistorted [2, 18], while the implementation of both OLS and GLS is not significantly more complex.

The results of the scalings and predictions on the ITPA02 data are presented in Tables 6 and 7, while the outcomes for the ITPA08 data can be found in Tables 8 and 9.

Table 4: Estimates of regression parameters in log-linear regression using the ITPA08 data, similar to Table 2.

Method		β_0	β_1	β_2	β_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	Av.	0.0478	0.73	0.796	0.952	53.7	89
	CI	± 0.0007	± 0.01	± 0.009	± 0.005	± 0.7	± 2
MAP	Av.	0.0491	0.69	0.83	0.926	50.8	82
	CI	± 0.0008	± 0.02	± 0.01	± 0.007	± 0.8	± 2
GLS	Av.	0.0484	0.75	0.79	0.954	53.7	90
	CI	± 0.0008	± 0.01	± 0.01	± 0.006	± 0.8	± 2

Table 5: Estimates of the observed standard deviations, in percentage, for log-linear GLS using the ITPA08 data, similar to Table 3.

	AUG	C-Mod	DIII-D	JET	JFT-2M	JT-60U
Av.	18	11.2	14.5	15.0	12.1	19
CI	± 1	± 0.5	± 0.6	± 0.3	± 0.4	± 2

Table 6: Estimates of regression parameters β_k and predictions for ITER, in nonlinear power-law regression on the original scale for the H-mode threshold power on the ITPA02 data set. The bootstrap averages are given, as well as the 95% credible intervals (CI).

Method		β_0	β_1	β_2	β_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	Av.	0.027	0.77	1.0	1.04	70	120
	CI	±0.008	±0.09	±0.1	±0.07	±20	±30
MAP	Av.	0.046	0.64	0.79	0.93	44	69
	CI	±0.004	±0.07	±0.08	±0.03	±4	±8
GLS	Av.	0.040	0.72	0.75	0.98	52	85
	CI	±0.004	±0.07	±0.08	±0.03	±4	±9

Table 7: Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the power threshold, expressed as percentage errors on P_{thr} itself, in the machines contributing to the ITPA02 data set, obtained using nonlinear power-law regression with GLS. The bootstrap averages are given, as well as the 95% credible intervals (CI).

	ASDEX	AUG	C-Mod	DIII-D	JET	JFT-2M	JT-60U	PBXM
Av. (%)	36	21	20	16	22	16	22	28
CI (%)	±9	±4	±3	±2	±4	±2	±5	±8

Table 8: Estimates of regression parameters in nonlinear power-law regression using the ITPA08 data, similar to Table 6.

Method		β_0	β_1	β_2	β_3	$\hat{P}_{\mathrm{thr},0.5}$ (MW)	$\hat{P}_{\mathrm{thr},1.0}$ (MW)
OLS	Av.	0.045	0.77	0.80	0.98	58	99
	CI	± 0.002	± 0.02	± 0.01	± 0.01	± 2	± 4
MAP	Av.	0.049	0.69	0.83	0.925	50.6	81
	CI	± 0.001	± 0.02	± 0.01	± 0.007	± 0.8	± 2
GLS	Av.	0.048	0.74	0.79	0.951	53.6	90
	CI	± 0.001	± 0.01	± 0.01	± 0.007	± 0.9	± 2

Table 9: Estimates of the observed standard deviations, in percentage, for nonlinear power-law GLS, similar to Table 7.

	AUG	C-Mod	DIII-D	JET	JFT-2M	JT-60U
Av.	17	11	15	14	12	18
CI	± 4	± 1	± 2	± 2	± 1	± 4

To obtain the tables for the observed standard deviations we again calculated relative errors. However, this time the relative errors are not the same for the measurements coming from a single machine, so we calculated an average for each machine (and similar for the credible interval).

5.3. Discussion

We now discuss the results of the experiments on real data, pointing out several differences between the regression results obtained by OLS, MAP and GLS, when applying these methods to different data sets and making use of different regression models.

We first consider the experiments based on log-linear scaling, from which we can obtain several noteworthy results:

- There are several instances, both in case of the 2002 and 2008 data sets, where the regression parameters estimated by OLS and, to some extent also MAP, differ significantly from those obtained by GLS. This is particularly the case for the dependence of the power threshold on density and surface area, as shown by the non-overlapping credible intervals.
- The parameters estimated by GLS are relatively similar for both data sets. Only the ITPA08 parameter for the density is just outside the credible interval of the corresponding ITPA02 parameter. A similar comment goes for MAP.
- The predictions for the power threshold are higher for the ITPA08 data than for the ITPA02 data. However, for GLS and MAP the difference is by far not as

pronounced as for OLS.

- In the case of ITPA08, the OLS parameters and predictions are very similar to those provided by GLS, while MAP slightly deviates from these results.
- The 95% credible intervals on the 2008 results are much narrower than those for the 2002 data set. This is the result of the improved conditioning of the 2008 data. It is also seen in the values of the $\sigma_{\mathrm{obs},\alpha}$, which are generally lower for the 2008 data.
- From the results in Table 3 for the ITPA02 data, we find an average observed relative error on P_{thr} across devices of 24.2%. The average modeled standard deviation, on the other hand, corresponds to an error bar of 16% on P_{thr}. This is somewhat higher than the average measurement error of 15% on P_{thr}, which is due to the additional uncertainty of the predictor variables propagating into the value of P_{thr} calculated from the scaling law. The important point, however, is that the average observed uncertainty (24.2%) is quite somewhat higher than the average modeled uncertainty (16%) (although still considerably lower than the upper bound of 38%, as calculated in Section 3). This is an indication of additional sources of uncertainty, on top of mere measurement error, causing the data points to deviate from the proposed regression model, as discussed already in Section 3. That extra uncertainty is detected by GLS, which, accordingly, raises the values of the observed standard deviations for each machine. This is the key to the enhanced flexibility and robustness of the GLS method. One also notices that, in the case of ASDEX, the observed variability around the scaling law is particularly high.
  On the other hand, from Table 5 follows an average observed error bar for the 2008 data of 15%. This should be compared to the average modeled error bar, which turns out to be 15% as well. Hence, in the case of the 2008 data, the observed data variability is, on average, the one expected due to measurement error. There is no need for GLS to augment the observed standard deviation over the modeled value. This also explains why on the ITPA08 data the three regression methods yield similar results.

When considering the nonlinear power-law scaling, we can additionally make the following interesting observations:

- In comparing the results of GLS between log-linear and nonlinear scaling and between the ITPA02 and ITPA08 data sets, again the good to excellent consistency of GLS can be noted. This is a solid argument in favor of the method. At the lower density level GLS gives predictions of P_{thr} of resp. 48 MW (log-linear ITPA02), 54 MW (log-linear ITPA08), 52 MW (nonlinear ITPA02) and 54 MW (nonlinear ITPA08), all of which are in the same range, particularly the latter three. This should be contrasted with the predictions by OLS at the same density, i.e. (in the same order) 38, 54, 70 and 58 MW. This indicates that the OLS predictions on the more recent ITPA08 database are more reliable than those on the ITPA02 data set, where OLS suffers from important inconsistencies. As far as MAP is concerned,

we obtain 46, 51, 44 and 51 MW . Hence, for MAP the consistency between the log-linear and nonlinear regression and between databases is clearly better than for OLS. On the other hand, for nonlinear regression the correspondence between the 2002 and 2008 data is worse for MAP (44 vs. 51 MW) than for GLS (52 vs. 54 MW). Again, this is because GLS has extra degrees of freedom, through the $\sigma_{\rm obs}$ parameters, to compensate for the additional uncertainty observed in the data, relative to what the model predicts.

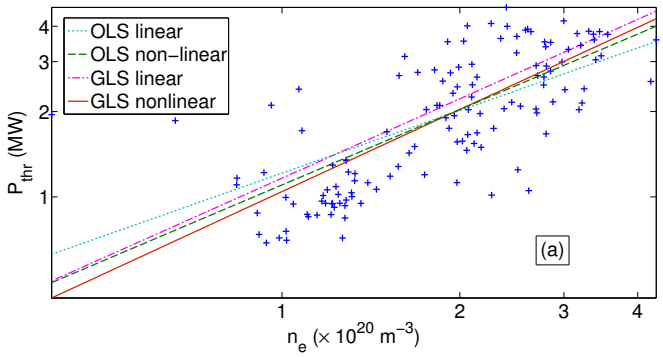
At higher densities the scatter on the predicted thresholds becomes more apparent, but still GLS yields comparable results in all cases.

- With nonlinear power-law regression using OLS, the 95% credible intervals are significantly wider than those provided by GLS and MAP.
- Still in the case of nonlinear OLS, the dependence on the magnetic field is considerably different for the two data sets. This leads to a power threshold predicted by OLS that is significantly higher for the 2002 data than for the 2008 data.

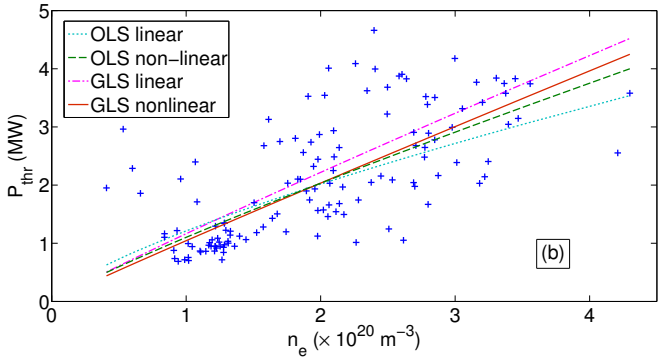
In order to further illustrate the improved estimates by GLS on the ITPA02 data, in comparison with OLS, we provide an example of a visual interpretation of the regression results as a function of density in Figure 6. The fits, obtained by log-linear OLS and GLS on the older ITPA02 data, are overlayed on a restricted data set from Alcator C-Mod at approximately constant magnetic field ($B_{\rm t} \approx 5.2$ T) and surface area ($S \approx 7.0$ m^2). Both Figure 6(a) and (b) contain the same data and fits, but (a) is drawn on the logarithmic scale, whereas (b) is on the original scale. In (a), OLS appears to be influenced more than GLS by the points on the upper left-hand side of the plot, which could be seen as data outliers, at least on the original scale. We have observed this trend also for many other subsets of the data. From Figure 6(b) it can be appreciated that even slight differences in the values of the regression coefficients can lead to relatively widely varying predictions of the power threshold in ITER.

We further wish to make a point regarding the commonly used visual assessment of the goodness-of-fit of a regression model. In Figure 7 the experimental power threshold is plotted against the one predicted by log-linear OLS and GLS using the older ITPA02 data. Although this figure does convey some information about the goodness of the fit, it has the disadvantage of suggesting somewhat misleadingly that OLS and GLS do not differ much in their predictions. Indeed, we have noted above that the regression coefficients estimated by GLS are quite different from those given by OLS, particularly in the density dependence, and the two methods predict significantly different power thresholds for ITER when applied to the less well conditioned ITPA02 data. Therefore, plots such as in Figure 7 are less suitable for comparing the performance of different regression methods, models or data sets.

Moreover, we do not mention a root mean square error or χ^2 value corresponding to the fit, since for GLS this would have to be based on geodesic distances, rendering a comparison in terms of such quantities with OLS and MAP meaningless.



(a)



(b)

Figure 6: Experimental threshold power versus density (ITPA02 data) with regression fits at constant field and surface area in Alcator C-Mod, on a logarithmic scale in (a) and original scale in (b).

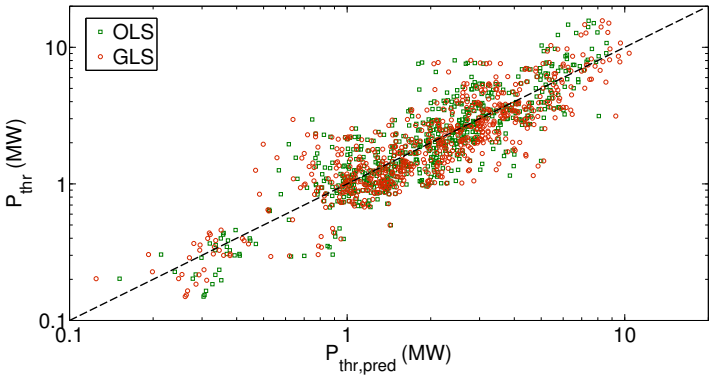


Figure 7: Experimental threshold power versus the power predicted by OLS and GLS regression for the ITPA02 data.

Table 10: Results of the Kadomtsev constraint (15) for OLS and GLS regression on the two data sets for log-transformed linear scaling.

Data set	ITPA02			ITPA08		
Method	OLS	MAP	GLS	OLS	MAP	GLS
$8\beta_1 + 5\beta_2 - 8\beta_3$	1.55	1.71	1.68	2.20	2.26	2.32

Table 11: Results of the Kadomtsev constraint (15) for OLS and GLS regression on the two data sets for power-law scaling.

Data set	ITPA02			ITPA08		
Method	OLS	MAP	GLS	OLS	MAP	GLS
$8\beta_1 + 5\beta_2 - 8\beta_3$	2.84	1.63	1.67	2.32	2.27	2.26

In addition, it should be noted that the above scalings were derived without additional constraints imposed by the physical system, other than those underlying the regression model. For instance, the Kadomtsev constraint regarding the dimensionality of the scaling is given by

$$8\beta_1 + 5\beta_2 - 8\beta_3 = 3. \tag{15}$$

From Tables 10 and 11, this is seen to be relatively well satisfied by our parameter estimates, particularly for the ITPA08 data. Alternatively, it would be possible to impose this constraint, or any other physics-based information, on the regression analysis, but we have not done this here.

Finally, although not a specific aim of the present paper, we can make a few comments about the attainability of the H-mode in ITER at different densities, given an available input power of 73 MW. First, one can note that OLS, MAP and GLS are close in their predictions of the power threshold, provided the latest version of the ITPA power threshold database is used. This is further confirmed by the results of GLS, which are relatively consistent across all experiments. The predictions also correspond to those of the currently used scaling law for the power threshold [5]. Looking at the predictions from the experiments, we may assume a threshold power of about 54 MW at a modest density of 0.5×10^{20} m^{-3}. Purely from this point of view the threshold should therefore be easily reachable at lower density. The situation is less clear at higher density (1.0×10^{20} m^{-3}), where the estimate by GLS of 90 MW may cause difficulties in reaching or maintaining the H-mode.

6. Conclusion

Several important scaling laws have been established in the past by means of statistical methods, providing essential design constraints for next-step fusion devices. With the

present paper we have aimed to show that a careful data-analytical study, combined with adoption of adequate and state-of-the-art techniques in probability theory, is mandatory in order to obtain reliable results from scaling laws. This is especially critical in the case of considerable uncertainty on the data, statistical models or physical models—circumstances that are rather common in fusion science.

The second goal of the paper was to present geodesic least squares (GLS) as a flexible and robust, yet easily implemented solution to model and data uncertainty in regression analysis. The essential difference with standard methods is that GLS is a non-Euclidean technique that carries out the regression analysis on a probabilistic manifold. It minimizes the difference (geodesic distance) between, on the one hand, the distribution of the dependent variable expected under the model (modeled distribution) and, on the other hand, the 'true' distribution of that variable, which relies as little as possible on the model assumptions (observed distribution). In this paper, we have described the simplest implementation of this idea for multilinear and power-law regression, leaving ample room for generalization and improvement of the method. For instance, GLS is not limited to Gaussian distributions, so the method can be readily transposed to other probabilistic manifolds.

Our experiments with synthetically generated data indicate that, in comparison with ordinary least squares and Bayesian maximum *a posteriori* estimation, GLS is considerably more robust against outliers and model uncertainty originating from a logarithmic transformation. In applying the log-linear and nonlinear regression analyses to fit the scaling law for the high-density branch of the L-H power threshold, using data from the ITPA 2002 and 2008 databases, consistent results were obtained by GLS. GLS was seen to be less affected by the validity of model assumptions, and by the quality and uncertainty of the data, as compared to standard OLS and, to some extent, even MAP.

In explaining the better performance of GLS compared to OLS and MAP, the flexibility offered by the observed distribution has proved to play a decisive role. In the present simple implementation of GLS, this role is essentially played by the observed standard deviation. Indeed, GLS allows the data uncertainty predicted by the model to be different from the empirically observed uncertainty, whereas with OLS and MAP they are identical by design. As a consequence, the degrees of freedom provided by the parameters of the regression model better serve their actual purpose: to parameterize a model that best describes a trend in the data, with minimal distraction by the data 'noise'.

Furthermore, although not demonstrated in the experiments in this paper, GLS regression has been shown to provide superior performance with respect to several other sophisticated methods [11]. This includes total least squares regression (TLS) [33], which is a typical errors-in-variables technique, and a robust method based on iteratively reweighted least squares (bisquare weighting) [34].

Although not a particular aim of the present paper, our case study for the L-H power threshold scaling law did confirm the validity of the scaling relation derived

earlier in [5]. In addition, GLS provides very similar results when applied to the older, less well conditioned database dating from 2002 [28]. This is an important motivation for pursuing scaling studies not only with a well-conditioned data set, but also using a state-of-the-art statistical methodology. In particular, application of linear regression analysis on log-transformed data assumed to follow a power law, is not recommended. Nevertheless, our experiments have pointed out that the data in the ITPA power threshold database from 2008 are sufficiently well conditioned to allow reliable results by means of simple OLS. On the other hand, it is clear that, in general, it can be dangerous to rely on the restrictive assumptions of OLS in regression studies.

We also wish to stress that regression analysis is of much more general use than for estimating scaling laws. Regression is routinely performed in fusion science for the purpose of model building and prediction in the context of physics studies. More often than not the assumptions underlying OLS are violated in fitting these models to data, and one has to revert to more powerful techniques. With the GLS method, we aim to provide a reliable tool to the fusion community for regression analysis in demanding circumstances (e.g. large uncertainties). For this purpose, future work will involve improving and generalizing GLS, particularly by reformulating the method in the framework of Bayesian probability theory on the Riemannian probabilistic manifold, yielding a full posterior distribution of the regression parameters and predictions. It should be emphasized that this is different from classic Bayesian methods, such as MAP, which operate in a flat Euclidean data space.

Finally, in the spirit of an ongoing tendency in fusion science, as in other disciplines, to aim for synergies between data-driven methods and physical understanding and techniques, we stress that it is perfectly possible to provide GLS regression with a set of constraints or, in the Bayesian framework, prior information regarding the underlying physics of the scaled quantity. This might be as simple as a set of rules encoding known relations between the quantities involved in the scaling, or it might involve incorporating a more detailed physical model into the regression model or in the prior information. This would allow taking into account the underlying physical mechanisms, in particular the physical picture of the L-H transition.

Acknowledgments

The authors wish to acknowledge the ITPA Topical Groups on Transport and Confinement and on Pedestal and Edge Physics for maintaining and kindly providing the data in the H-mode threshold databases.

References

[1] Doyle E *et al.* 2007 *Nucl. Fusion* **47** S18–S127
[2] McDonald D *et al.* 2006 *Plasma Phys. Control. Fusion* **48** A439–A447
[3] Murari A *et al.* 2012 *Nucl. Fusion* **52** 063016
[4] Murari A *et al.* 2013 *Nucl. Fusion* **53** 043001

[5] Martin Y *et al.* 2008 *J. Phys: Conf. Ser.* **123** 012033
[6] Sivia D and Skilling J 2006 *Data Analysis: a Bayesian Tutorial* 2nd ed (Oxford: Oxford University Press)
[7] von der Linden W, Dose V and von Toussaint U 2014 *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge: Cambridge University Press)
[8] Verdoolaege G, Fischer R, Van Oost G and JET-EFDA Contributors 2010 *IEEE Trans. Plasma Sci.* **38** 3168–3196
[9] Verdoolaege G 2015 *Entropy* **17** 4602–4626
[10] Verdoolaege G *et al.* 2012 *Plasma Phys. Control. Fusion* **54** 124006
[11] Verdoolaege G 2014 *Rev. Sci. Instrum.* **85** 11E810
[12] Verdoolaege G 2013 Geodesic least squares regression on information manifolds *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (*AIP Conference Proceedings* vol 1636) (Meville, NY) pp 43–48
[13] Verdoolaege G 2014 Geodesic least squares regression for scaling studies in magnetic confinement fusion *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (*AIP Conference Proceedings* vol 1641) (Melville, NY) pp 564–571
[14] Fuller W 2006 *Measurement Error Models* (New York: John Wiley & Sons, Inc.)
[15] Jaynes E 1990 Straight line fitting—a Bayesian solution Presented at the 10$^{\text{th}}$ International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Unfinished manuscript available at http://bayes.wustl.edu/etj/articles/leapz.pdf
[16] Preuss R and Dose V 2005 Errors in all variables *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (*AIP Conference Proceedings* vol 803) (Melville, NY) pp 448–455
[17] von Toussaint U, Frey M and Gori S 2009 Fitting of functions with uncertainties in dependent and independent variables *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (*AIP Conference Proceedings* vol 1193) (Melville, NY) pp 302–310
[18] Xiao X *et al.* 2011 *Ecology* **92** 1887–1894
[19] Verdoolaege G, Karagounis G, Murari A, Vega J, Van Oost G and JET-EFDA Contributors 2012 *Fusion Sci. Technol.* **62** 356–365
[20] Verdoolaege G and Scheunders P 2011 *J. Math. Imaging Vis.* **43** 180–193
[21] Amari S and Nagaoka H 2000 *Methods of Information Geometry* (New York: American Mathematical Society)
[22] Burbea J and Rao C 1982 *J. Multivariate Anal.* **12** 575–596
[23] Čenkov N 1982 *Statistical decision rules and optimal inference* (*Translations of Mathematical Monographs* vol 53) (Providence, RI: American Mathematical Society)
[24] Rao C 1987 *Differential Geometry in Statistical Inference* (Hayward, CA: Institute of Mathematical Statistics) chap Differential Metrics in Probability Spaces
[25] Beran R 1977 *Ann. Stat.* **5** 445–463
[26] Pak R 1996 *Stat. Probab. Lett.* **26** 263–269
[27] Gill P, Murray W and Wright M 1991 *Numerical linear algebra and optimization, Vol. 1* (Boston, MA: Addison Wesley)
[28] Snipes J *et al.* 2002 Multi-machine global confinement and H-mode threshold analysis *Proceedings of the 19*$^{\text{th}}$ *IAEA Fusion Energy Conference* CT/P-04 (Lyon, France)
[29] Casella G and Berger R 2002 *Statistical Inference* 2nd ed (Hampshire, UK: Cengage Learning)
[30] Ryter F and the H-Mode Database Working Group 1996 *Nucl. Fusion* **36** 1217–1264
[31] Ryter F and the H-Mode Threshold Database Group 2002 *Plasma Phys. Control. Fusion* **44** A415–A421
[32] 2008 http://efdasql.ipp.mpg.de/threshold
[33] Markovsky I and Van Huffel S 2007 *Signal Process.* **87** 2283–2302
[34] Maronna R, Martin D and Yohai V 2006 *Robust Statistics: Theory and Methods* (New York: Wiley)